A Network Model for Wide Area Access to Structural Information

D.J.Dailey*, K.Eno**, G.L. Zick* and J.F. Brinkley**
Depts. Electrical Engineering* and Biological Structure**
University of Washington, Seattle, WA 98195

Structural information in medicine is information about the physical body. Recent advances in medical imaging and biotechnology have greatly increased the amount and importance of structural information, and advances in networking envisioned by the High Performance Computing and Communication Initiative (HPCC) will allow this kind of information to be delivered to remote clients over wide area networks. One of the most important factors determining the usability of such a client-server configuration is the time delay between the request for information from the server, and its presention to the user at the client. In this paper we present a model for predicting the performance of a structural information client based on the ping time, a simple, unobtrusive network measurement. Preliminary results suggest that the relationship between ping time and transfer time for large files is linear, which if borne out by more data, will allow the performance of structural information clients at remote sites to be predicted without the expense of installing them first. At the same time, such a model will be useful for planning improvements to the network in those sites which could most benefit by wide area access to structural information.

INTRODUCTION

One of the more important kinds of medical information is structural information about the physical organization of the body, at levels ranging from gross anatomy to molecules. Examples of structural information include molecular sequences, medical images, 3-D graphics reconstructions, anatomic terminology, and semantic relationships. Although structural information has always been important in medicine, recent advances in imaging and biotechnology have greatly increased not only the amount, but also the importance of this kind of information [1].

Two characteristics of structural information that distinguish it from most medical information are that it is multimedia and it is large. These characteristics have until recently prevented widespread digital access to structural information, but the information superhighways envisioned by the High Performance Computing and Communication Initiative (HPCC) [2] will allow this kind of information to be transmitted rapidly enough to permit widespread sharing of structural information. Since much of medicine is structural in nature this capability should lead to many applications.

In the Digital Anatomist Program at the University of Washington we have developed one such application for wide area access to structural information, the Digital Anatomist Browser [3, 4]. In this application images and other structural information are stored on a server computer, and a client program, running on a Macintosh, accesses the server over the network to deliver the information to end users. The browser is currently used by medical students for anatomy review, and a demonstration of the potential of such a configuration for distance learning in medicine is currently in progress with a client at the National Library of Medicine, and the server in Seattle.

The usability of such a remote learning system depends to a large extent on user response time, which, in the absence of high speed networks, becomes a dominant factor as physical distance between client and server increases. In this paper we develop a network model that, given several easily measured parameters, allows one to predict the expected performance of a particular client-server configuration, thereby determining whether the performance is likely to be good enough to warrant installation of the client. Multiple such predictions could also be used to plan future improvements to the network in areas of need.

PERFORMANCE FACTORS

There are three major factors that contribute to the response time of a client-server configuration for structural information retrieval. These are the delay at the server, the delay at the client, and the network delay.

The delays at the client and server end we label "static delays". These delays are independent of the network and are generally associated with the particular instantiation of the client and server. Examples of things that effect static delays include the type of CPU, the performance of the disk drives, and the amount of memory in both the client and the server. While these parameters will vary depending on the platform, once selected they are fixed.

The network delays we label "dynamic delays". These are associated with the location of the client and servers on the network and level of the network congestion. It this later type of delay that this paper seeks to quantify. Later work will combine these with static delays associated with a particular client and server to predict the actual performance of the client at remote sites, without the necessity of actually installing the client at the site.

NETWORK DELAY MODEL

The goal of our network model is to predict the performance of the Digital Anatomist Browser or a similar structural information retrieval client at an arbitrary location on the Internet. To do this two measurements are necessary; first, a metric to measure the "network distance" between the client and the server and second, a measure of the data transfer delay expected as a function of network distance.

To quantify the network distance between the client and the data server we use the round trip time for a packet of data to travel from the client to the server and back. These measurements use the Internet Control Message Protocol (ICMP) implemented in the Unix *ping* command. This metric quantifies network distance without knowledge of the number and type of links between the sites. The use of this metric provides a model that is applicable without detailed knowledge of the physical connectivity of the network. This independence from the physical design of the network is important because the path, through the network, from the server to the client may change as the Internet evolves.

The Digital Anatomist Browser transfers images and other data from remote servers over the Internet. This transfer is one of the key steps in the operation of the Browser, and large delays in this step can dramatically change the user response time and thus adversely effect usability. The image data can be from 200 kilobytes to 1 megabyte in size which, by todays standards, is a significant amount of data to transfer in a real time application. In this work we estimate the time for the data transfer using the File Transfer Protocol as implemented by the Unix *ftp* command. This command provides a transfer time estimate for each transfer. Just as our network distance metric (ping time) is independent of the physical network design this metric for the delay introduced by the data transfer can be obtained without detailed knowledge of the physical network. We use these two metrics to develop a model for the network delays expected at arbitrary locations on the Internet.

The model we use to predict network performance is based on the hypothesis that there is a relationship between the network distance, as measured by the *ping* command (labeled x), and the time it takes to move structural data, as measured by the time to transfer a fixed size file using the *ftp* command (labeled y).

The general form for such a model is:

$$y = f(x; a, b). \tag{1}$$

We further hypothesize that this relationship is linear. The linear model is written

$$y = ax + b. \tag{2}$$

Unlike standard linear regression problems where there is a control variable (e.g. x) and an observed variable (e.g. y) this model has two observables (e.g. x, y) and no control variable.

Further, the network delay associated with each remote site has an independent set of statistics associated with the measurements specific to that site. We quantify the statistics of the *ftp* and *ping* observations at each site in terms of the mean and standard deviation of the measurements to that site. Hence the actual observables are:

1. the mean (x_i) and standard deviation (σ_{x_i}) of the roundtrip ping time for the ith of N remote sites
2. the mean (y_i) and standard deviation (σ_{y_i}) of the file transfer time for each of the sites.

For M observations at each site the unbiased estimate of these values from measured data are:

$$x_i = \frac{1}{M}\sum_{j=1}^{M} x_{ij}, \qquad y_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}, \qquad (3)$$

$$\begin{aligned} \sigma_{x_i} &= \frac{1}{M-1}\sum_{j=1}^{M}(x_{ij} - x_i)^2, \\ \sigma_{y_i} &= \frac{1}{M-1}\sum_{j=1}^{M}(y_{ij} - y_i)^2, \end{aligned}$$

where x_{ij} and y_{ij} are the jth measurements from the ith site. These estimates of the statistics at each site are used to obtain a maximum likelihood estimate of the parameters (a, b) from equation 2. The maximum likelihood methodology is chosen over the simpler linear regression based on two observations; 1) the observables used in estimating the parameters (a, b) are statistical quantities and 2) the statistics at each site are independent of the other sites. These two conditions violate the assumptions necessary to use a simple regression technique. The following development provides a maximum likelihood methodology for estimating the parameters of our model in the context of our observables.

The deviation of a measurement from the linear model is written:

$$Error: \quad \epsilon_i = y_i - y(x_i; a, b). \qquad (4)$$

We approximate these errors as being normally distributed so that the probability density function can be written:

$$f(\epsilon_i; x_i, y_i, a, b) = \frac{1}{2\pi\sigma_i} exp\left\{-\frac{1}{2}\left(\frac{y_i - y(x_i; a, b)}{\sigma_i}\right)^2\right\} \qquad (5)$$

where σ_i is the standard deviation of the errors ϵ_i. Now if x_i and y_i are uncorrelated in the cluster of measurements for the ith site,

$$\sigma_i^2 = \sigma_{y_i}^2 + a^2\sigma_{x_i}^2 \qquad (6)$$

and substituting 6 into equation 5 we get.

$$f(\epsilon_i) = \frac{1}{2\pi(\sigma_{y_i}^2 + a^2\sigma_{x_i}^2)^{1/2}} \times exp\left\{-(\frac{1}{2})\frac{(y_i - ax_i - b)^2}{\sigma_{y_i}^2 + a^2\sigma_{x_i}^2}\right\}. \qquad (7)$$

A likelihood function for these statistics is written,

$$L(\epsilon; a, b) = \prod_{i=1}^{N} f(\epsilon_i; a, b). \qquad (8)$$

Maximizing this likelihood function is equivalent to minimizing the negative of its log with respect to the parameters, and so the best estimate of the parameters satisfies

$$\begin{aligned} \frac{\partial}{\partial a}(-ln(L(\epsilon; a, b)) &= 0 \qquad (9) \\ \frac{\partial}{\partial b}(-ln(L(\epsilon; a, b)) &= 0. \qquad (10) \end{aligned}$$

This results in coupled nonlinear equations. The intercept term (b) is written,

$$b = \frac{\sum_{i=1}^{N} \frac{y_i - ax_i}{\sigma_{y_i}^2 + a^2\sigma_{x_i}^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_{y_i}^2 + a^2\sigma_{x_i}^2}}. \qquad (11)$$

The expression resulting from equation 9:

$$\sum_{i=1}^{N} \frac{a\sigma_{x_i}}{A} = \sum_{i=1}^{N} \frac{2ABx_i + 2aB^2\sigma_{x_i}^2}{A^2} \qquad (12)$$

where,

$$\begin{aligned} A &= \sigma_y^2 + a^2\sigma_{x_i}^2 \\ B &= y_i - ax_i - b \end{aligned}$$

does not result in a closed form expression for the coefficient a. Equations 11 and 12 are solved iteratively for the parameters a and b. This is done using an interval halving technique for which the starting values are obtained from a linear regression of the mean *ftp* value on the mean *ping* value from each site.

MODEL PARAMETERIZATION

To use the method just developed a number of measurements of both *ping* times and *ftp* times, at varying network distances, are needed. With the assistance of a number of Internet sites these statistics were gathered. To estimate means and standard deviations for each site the *ping* time was measured 1000 times, and the *ftp* delay was measured 100 times. Both commands used data packets of 590 bytes in length and the file transferred by *ftp* was 1003520 bytes in total length. The result of fitting the model from equation 2 to the Internet data is shown in figure 1. The confidence interval shown was constructed using some assumptions about future observations.

We first identify a variable associated with the deviation from the predicted model,

$$z(x_i) = \frac{y_i - y(x_i)}{\sigma_i} \tag{13}$$

which is a normal random variable with the $\sigma_z = 1$ and mean of zero. Thus

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{14}$$

To find the probability β that a value is in the range $-b$ to b we integrate,

$$\beta = P[z(x_i) \leq b] \tag{15}$$

$$= \int_{-b}^{b} p(z)dz \tag{16}$$

$$= \frac{2}{\sqrt{2}} erf(b/2) \tag{17}$$

where $erf(*)$ is the error function. The relationship between the confidence probability β and the bound b is

$$b = 2erf^{-1}(\frac{\sqrt{2}}{2}\beta). \tag{18}$$

Recalling our definition of z

$$P\left[\frac{y_i - y(x_i)}{\sigma_i} \leq b\right] = \beta \tag{19}$$

which provides us a confidence interval expression,

$$P\left[y_i \leq b\sigma_i + y(x_i)\right] = \beta \tag{20}$$

and using

$$\sigma^2 \approx \frac{1}{N-1}\sum_{i=1}^{N}(y_i - y(x_i))^2 \tag{21}$$

we can estimate the confidence intervals. For example, the 98% confidence interval is obtained by setting

$\beta = 0.98$ in equation 18 and calculating the bound b. This value is then used with the estimated σ and the model to place a confidence bound on figure 1.

DISCUSSION

Figure 1 suggests that there is indeed a linear relationship between ping time and file transfer time, although more data points must be obtained before the relationship can be demonstrated with any certainty. To our knowledge a linear relationship of this type has never been actually demonstrated. We believe such a demonstration has not been presented previously because there are few if any applications such as the Digital Anatomist Browser that interactively transfer structural data over the network. However, just as many text-based remote database servers are now appearing [5, 6], the HPCC will lead to structural data servers as well.

The model we have developed here, when made more robust with a larger number of tests to more remote sites, will allow us to measure a network distance in an unobtrusive, easy way and to predict, with a specified confidence level, the dynamic delays expected in remote operation of a structural information client. When the predicted dynamic delays are combined with measured static delays for a given client and server configuration it will be possible to predict the performance of the client before actually installing it at the remote site. If the predicted performance is too poor much installation effort can be saved, and instead plans can be made for upgrading the network.

References

[1] Brinkley, J.F. Structural informatics and its applications in medicine and biology. Academic Medicine 66:589-591; 1991.

[2] High Performance Computing Act of 1991. reprinted in Electronic Networking 2:7-9: 1992.

[3] Eno, K.; Sundsten. J.W.; Brinkley, J.F. A multimedia anatomy browser incorporating a knowledge base and 3-D images. Proc. Symposium on Computer Applications in Medical Care. 15:727;1991.

[4] Brinkley, J.F.: Eno. K.; Sundsten. J.W. Knowledge-based client-server approach to structural information retrieval: the Digital Anatomist Browser. Comp. Meth. Prog. Biomed. In Press: 1993.

[5] B. Kahle, H. Morris. F. Davis, K. Tiene. C. Hart and R. Palmer, Wide area information servers: an executive information system for unstructured files. Electronic Networking 2:59-68; 1992.

[6] T. Berners-Lee, R. Cailliau, J-F Groff and B. Pollerman, World-wide web: the information universe, Electronic Networking 2:52-58: 1992.

Acknowledgements

This research was supported by National Library of Medicine contract NO1-LM-1-3506. We would also like to thank the following internet sites for their willingness allow us access to their machines: Washington University, Columbia University, Stanford University, Simon Fraser University, and University of Central Florida.

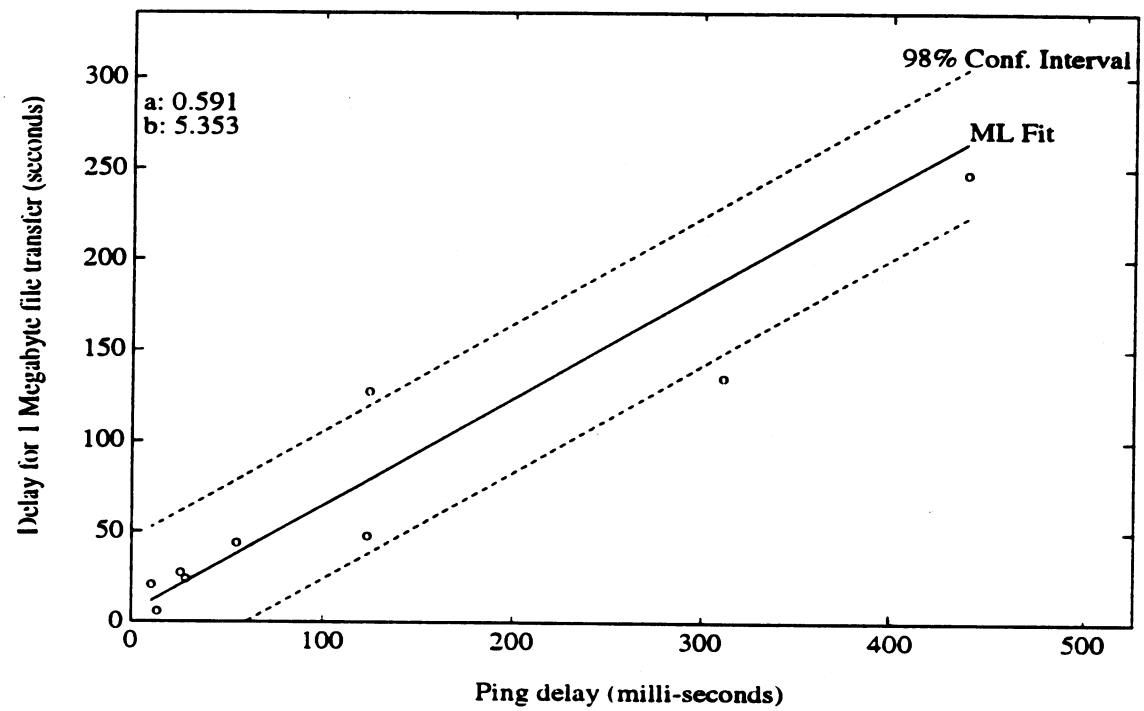

Figure 1: FTP transfer time as a function of ping distance.